# Identification and characterization of essential genes in the human genome

**Tim Wang,**[1,2,3,4] **Kıvanç Birsoy,**[1,2,3,4]* **Nicholas W. Hughes,**[3] **Kevin M. Krupczak,**[2,3,4] **Yorick Post,**[2,3,4] **Jenny J. Wei,**[1,2] **Eric S. Lander,**[1,3,5]†‡ **David M. Sabatini**[1,2,3,4,6]†‡

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02139, USA. [2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. [3]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA. [4]David H. Koch Institute for Integrative Cancer Research at MIT, Cambridge, MA 02139, USA. [5]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [6]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*Present address: Laboratory of Metabolic Regulation and Genetics, The Rockefeller University, New York, NY 10065, USA.

†These authors contributed equally to this work.

‡Corresponding author. E-mail: lander@broadinstitute.org (E.S.L.); sabatini@wi.mit.edu (D.M.S.)

**Large-scale genetic analysis of lethal phenotypes has elucidated the molecular underpinnings of many biological processes. Using the bacterial clustered regularly interspaced short palindromic repeats (CRISPR) system, we constructed a genome-wide single-guide RNA (sgRNA) library to screen for genes required for proliferation and survival in a human cancer cell line. Our screen revealed the set of cell-essential genes, which was validated by an orthogonal gene-trap-based screen and comparison with yeast gene knockouts. This set is enriched for genes that encode components of fundamental pathways, are expressed at high levels, and contain few inactivating polymorphisms in the human population. We also uncovered a large group of uncharacterized genes involved in RNA processing, a number of whose products localize to the nucleolus. Lastly, screens in additional cell lines showed a high degree of overlap in gene essentiality, but also revealed differences specific to each cell line and cancer type that reflect the developmental origin, oncogenic drivers, paralogous gene expression pattern, and chromosomal structure of each line. These results demonstrate the power of CRISPR-based screens and suggest a general strategy for identifying liabilities in cancer cells.**

The systematic identification of essential genes in microorganisms has provided critical insights into the molecular basis of many biological processes (*1*). Similar studies in human cells have been hindered by the lack of suitable tools. Moreover, little is known about how the set of cell essential genes differs across cell types and genotypes. Differentially essential genes are likely to encode tissue-specific modulators of key cellular processes and important targets for cancer therapies. Here, we used two independent approaches for inactivating genes at the DNA level to define the cell-essential genes of the human genome.

The first approach employs the CRISPR/Cas9-based gene editing system, which has emerged as a powerful tool to engineer the genomes of cultured cells and whole organisms (*2*, *3*). We and others have shown that lentiviral (single-guide RNA) sgRNA libraries can enable pooled loss-of-function screens and have used the technology to uncover mediators of drug resistance and pathogen toxicity (*4–6*). To systematically identify cell-essential genes, we constructed a novel library, which was optimized for high cleavage activity, and performed a proliferation-based screen in the near-haploid human KBM7 chronic myelogenous leukemia (CML) cell line (Fig. 1 and supplementary text S1).

The unusual karyotype of these cells also allows for an independent method of genetic screening. In this approach, null mutants are generated at random through retroviral gene-trap mutagenesis, selected for a phenotype, and monitored by sequencing the viral integration sites to pinpoint the causal genes (*7*). Positive selection-based screens using this method have identified genes underlying processes such as epigenetic silencing and viral infection (*7–9*). We extended this technique by developing a strategy for negative selection and conducted a screen for cell-essential genes (Fig. 1 and supplementary text S2).

For both methods, we computed a score for each gene that reflects the fitness cost imposed by inactivation of the gene. We defined the CRISPR score (CS) as the average $\log_2$ fold-change in the abundance of all sgRNAs targeting a given gene, with replicate experiments showing a high degree of reproducibility ($r = 0.90$) (Fig. 2A, fig. S1A, and table S2). Of the 18,166 genes targeted by the library, 1,878 scored as essential for optimal proliferation in our screen, although this precise number depends on the cutoff chosen (Fig. 2A and tables S2 and S3). Overall, this fraction represents ~10% of genes within our dataset or roughly 9.2% of the entire genome (many of the genes not targeted by our library encode olfactory receptors that are unlikely to be cell-essential). Gene products that act in a non-cell-autonomous manner are not expected to score as essential in this pooled setting (fig. S1B).

We defined the gene-trap score (GTS) as the fraction of insertions in a given gene occurring in the inactivating orientation. Because the accuracy of this score depends on the

depth of insertional coverage, we set a requirement on the minimum number ($n = 65$) of anti-sense inserts in a gene needed for inclusion in our analysis by measuring the concordance between replicate experiments (Fig. 2B; fig. S1, C and D; table S4; and supplementary text S2). For the 7,370 genes on the haploid chromosomes that exceeded this threshold, the GTS was well-correlated with the CS and with results from a co-published study employing a similar gene-trap approach ($r = 0.68$) (Fig. 2C; fig. S1, E and F; and supplementary text S3). The strong correspondence between the overlapping sets of cell-essential genes defined by the two methods provides support for the accuracy of the CRISPR scores for the full set of 18,166 genes.

The two methods differed with respect to the diploid chromosome 8. Whereas the gene-trap screen failed to detect any cell-essential genes on this chromosome, the CRISPR screen uncovered a similar proportion of cell-essential genes on all the autosomes (Fig. 2, A and C). These observations indicate that (i) the vast majority of cell-essential genes are haplosufficient and (ii) biallelic inactivation occurs at high frequency in our CRISPR screen (4).

To assess the accuracy of our scores with other measures of gene essentiality, we relied on functional profiling experiments conducted in yeast *S. cerevisiae* as a benchmark (1, 10). Specifically, we ranked genes common to all datasets by their scores in each dataset –CRISPR, gene trap, scores from similar loss-of-function RNA interference (RNAi) screens (11), and, as a naïve proxy for gene essentiality, gene expression levels determined by RNA-seq– and compared these rankings with the essentiality of yeast homologs. The CRISPR and gene-trap methods had significantly stronger correlations with the yeast results than did the RNAi screens or gene expression, which performed similarly to each other (both methods: $P < 10^{-4}$, permutation test) (Fig. 2D). Based on additional comparisons with yeast gene essentiality, we also found that (i) our new optimized sgRNA library gave better results than those from screens using older unoptimized libraries (4, 5) and (ii) the coverage of this library (~10 constructs per gene) approaches saturation, as evidenced by down-sampling (decreasing the coverage by randomly eliminating subsets of data) (fig. S2, A and B). Together, our results suggest that scores from the CRISPR and gene-trap screens both provide accurate measures of the cell-essentiality of human genes.

Essential genes should be under strong purifying selection and should thus show greater evolutionary constraint than non-essential genes (12). Consistent with this expectation, the essential genes found in our screens were more broadly retained across species, showed higher levels of conservation between closely related species, and contain fewer inactivating polymorphisms within the human species, as compared to their dispensable counterparts (Fig. 2, E and G). Essential genes also tend to have higher expression and encode proteins that engage in more protein-protein interactions (13–15). These patterns were also ob-

served in our CRISPR dataset (Fig. 2, H and I).

In *S. cerevisiae*, genes with paralogous copies in the genome show a lower degree of essentiality –due presumably to at least partial functional overlap (16). Surprisingly, meta-analysis of knockout mouse collections has suggested that there is no such correlation in mammals (17, 18). However, others have challenged this interpretation because the genes analyzed were far from a random sample (19). Using the results from our genome-wide screens, we revisited this question and observed that genes with paralogs are indeed less likely to be essential, consistent with the idea that paralogs can provide functional redundancy at the cellular level (Fig. 2J).

To examine the functions of the cell-essential genes we used gene set enrichment analysis (GSEA) and found strong enrichment for many fundamental biological processes, such as DNA replication, RNA transcription and mRNA translation (fig. S3A) (20). Whereas most of the genes could be assigned to such well-defined pathways, no function has been ascribed to about 330 of the cell-essential genes (18%) (Fig. 3A). For this set of uncharacterized genes, an analysis of the domains within their encoded gene products and comparisons to proteomic datasets from organellar purifications revealed substantial enrichment in proteins found in the nucleolus and those containing domains associated with RNA processing (fig. S3, B and C) (21).

We characterized three such genes, *C16orf80*, *C3orf17*, and *C9orf114*, whose mRNA expression patterns across the cancer cell line encyclopedia (CCLE) were correlated with that of genes involved in RNA processing (Fig. 3B). We validated the essentiality of these genes in short-term proliferation assays and detected localization of their products to the nucleus (C16orf80) or nucleolus (C3orf17 and C9orf114) (Fig. 3, C and D) (22). Additionally, mass spectrometric analyses of anti-FLAG-immunoprecipitates prepared from KBM7 cells expressing FLAG-tagged C16orf80, C3orf17, and C9orf114 revealed interactions with multiple subunits of the spliceosome, ribonuclease (RNAse) P/MRP, and H/ACA small nucleolar ribonucleoprotein (snoRNP) complexes, respectively (Fig. 3E). These results implicate *C16orf80* in splicing, consistent with its association with mRNAs; *C3orf17* in rRNA/tRNA processing; and *C9orf114* in RNA modification (23). More broadly, our results indicate that the molecular components of many critical cellular processes, especially RNA processing, have yet to be fully defined in mammalian cells.

To determine how the set of essential genes differs among cell lines, we screened another CML cell line (K562) and two Burkitt's lymphoma cell lines (Raji and Jiyoye) using the CRISPR system (tables S2 and S3). Overall, the sets of essential genes in the four cell lines showed a high degree of overlap (Fig. 4A). Importantly, out of these four cell lines, the KBM7 CRISPR results showed the highest correlation with the KBM7 gene-trap dataset, suggesting that the few differences observed are likely to be biologically meaningful

(fig. S4A).

We focused first on genes found to be essential in only one of the four cell lines. The Raji, Jiyoye, and KBM7 cell lines had 6, 7, and 19 such genes, respectively (fig. S4B and table S5). One example was *DDX3Y*, which resides in the non-pseudoautosomal region of the Y chromosome and was required only in Raji cells (Fig. 4B). Its X-linked paralog, *DDX3X*, was essential in KBM7 and K562 cells (Fig. 4E). Both genes encode DEAD-box helicases that likely have similar cellular functions (*24*). Thus, the dependence on one paralog might reflect functional absence of the other paralog. Indeed, DNA sequencing of *DDX3X* in Raji cells revealed hemizygous mutations in the 5'-splice site of intron 8 that resulted in the production of a truncated mRNA transcript (Fig. 4, C and D). Conversely, *DDX3Y* was not expressed in KBM7 cells and was not present in K562 cells, which are of female origin (Fig. 4E). Introduction of wild-type *DDX3X* cDNA into Raji cells fully rescued the proliferation defect resulting from *DDX3Y* loss, indicating that the paralogous genes are essential and functionally overlapping (Fig. 4F). Essential paralogous gene pairs, involved in glucose metabolism (*HK1/2* and *SLC2A1/3*) and cell-cycle regulation (*CDK4/6*), were also observed in the Jiyoye line (fig, S4C and supplementary text S4). Vulnerabilities due to the loss of a paralogous partner may serve as targets for highly personalized anti-tumor therapies (*25*).

In some cases, cell line-specific essentiality of paralogous genes did not reflect differential expression. For example, the transcription factors *GATA1* and *GATA2* are expressed in both K562 and KBM7 cells, but the first is specifically essential in K562 cells and the second in KBM7 cells (fig. S4D). These master regulators are known to promote proliferation and survival during distinct developmental stages in the hematopoietic lineage – *GATA1* is required for the survival of erythroid progenitors and *GATA2* for the maintenance and proliferation of immature hematopoietic progenitors (*26*). These two cell types likely correspond to the cells-of-origin of the two CML lines (*27, 28*). We also identified similar instances of genes required for cell line-specific functions such as NF-κB pathway regulation and homology-directed DNA repair in Raji and KBM7 cells, respectively (fig. S4, E and F, and supplementary text S5).

Whereas the other three cell lines showed only a few cell line-specific essential genes, the K562 cell line was an outlier, with 63 such genes (fig. S4B). Oddly, these genes showed no discernible common functions and many were not even expressed in the K562 cell line. (Additionally, a few encoded secreted factors whose loss would not be expected to be lethal in a pooled screen.) This mystery was resolved when we examined the chromosomal location of the genes: strikingly, the majority (39 of the 63 genes) reside near 9q34 and 22q11. These two regions are translocated to produce the *BCR-ABL* oncogene and are present in a high-copy tandem amplification in K562 cells (fig. S5, A and B) (*29*). Notably, all 61 contiguous genes within the amplicon on 22q11 scored as essential, suggesting that sgRNA-mediated cleavage in this repeated region induces cytotoxicity in a manner unrelated to the function of the target genes themselves (Fig. 4G and fig. S5F). Indeed, two sgRNAs targeting non-genic sites within the amplicon were toxic to K562, but not KBM7 cells. They increased the abundance of phosphorylated histone H2AX (γH2AX), a marker of DNA damage, and induced erythroid differentiation, which occurs upon DNA damage in this cell line (Fig. 4, H to J, and fig. S5C) (*30*). We obtained similar results in another cell line, HEL, which contains a highly amplified region surrounding the *JAK2* tyrosine kinase (fig. S5, D and E). Together, these findings indicate that lethality upon Cas9-mediated cutting may also reflect chromosome structure and therefore should be evaluated in light of copy-number information.

Finally, we looked for consistent differences in essential genes between the two CML and two Burkitt's lymphoma lines. Such genes might represent attractive targets for antineoplastic therapies as their inhibition is less likely to be broadly cytotoxic. Overall, we identified 33 genes that were specifically essential in the CML lines and 15 genes in the Burkitt's lymphoma lines (Fig. 4K and table S6). As a control, permuted comparisons – that is, a set containing of one CML and one Burkitt's line vs. the complementary sets – showed roughly half as many "set-specific" essential genes (fig. S6, A to C).

In the CML lines, the top two genes were *BCR* and *ABL1*, consistent with the known essentiality of the BCR-ABL translocation product and the therapeutic effect of BCR-ABL inhibitors such as imatinib (*31*). Additional members of the BCR-ABL signaling pathway, son of sevenless homolog 1 (*SOS1*), growth factor receptor-bound protein (*GRB2*), and GRB2-associated binding protein 2 (*GAB2*) scored strongly as well (ranked 3, 4, and 7, respectively). Network analysis of the other top hits also uncovered several genes encoding assembly factors for the electron transport chain, as well as enzymes involved in folate-mediated one-carbon metabolism. These results suggest additional potential targets for CML therapy (Table S6).

In the B cell-derived Burkitt's lymphoma cell lines, the top genes included three B cell lineage transcription factors early B cell factor 1 (*EBF1*), POU class 2 associating factor 1 (*POU2AF1*), and paired box 5 (*PAX5*) (ranked 3, 6, and 8, respectively). Each of these genes is the target of recurrent translocations in lymphoma (*32–34*). Enhancers of the corresponding three gene loci all show a high level of bromodomain containing 4 (BRD4) occupancy in Ly1 cells, a related diffuse large B cell lymphoma cell line, suggesting bromodomain inhibitors such as JQ1 as potential treatments (*35*). Other selectively essential genes included *MEF2B*, a transcriptional activator of *BCL6*, B cell lymphoma 6, and *CCND3*, cyclin D3, both of which are frequently mutated and implicated in the pathogenesis of various lymphomas (*36*). Intriguingly, the top two hits, *CHM* and *RPP25L,* do not appear to have specific roles in B cells; rather, their dif-

ferential essentiality is likely explained by the lack of expression of their paralogs, *CHML* and *RPP25*, in both of the Burkitt's lymphoma cell lines studied (fig. S6D).

In conclusion, we used two complementary and concordant approaches, CRISPR and gene-trap, to define the cell-essential genes in the human genome. Although the gene-trap method is suitable only for loss-of-function screening in rare haploid cell lines, the CRISPR method is broadly applicable. Extending our analysis across different cell lines and tumor types, we developed a framework to assess differential gene essentiality and identify potential drivers of the malignant state. The method can be readily applied to more cell lines per cancer type to eliminate idiosyncrasies particular to a given cell line and to more cancer types to systematically uncover tumor-specific liabilities that might be exploited for targeted therapies.

## REFERENCES AND NOTES

1. G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, M. Johnston, Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387–391 (2002). Medline doi:10.1038/nature00935
2. L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, F. Zhang, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013). Medline doi:10.1126/science.1231143
3. H. Wang, H. Yang, C. S. Shivalila, M. M. Dawlaty, A. W. Cheng, F. Zhang, R. Jaenisch, One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013). Medline doi:10.1016/j.cell.2013.04.025
4. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014). Medline doi:10.1126/science.1246981
5. O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, F. Zhang, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014). Medline doi:10.1126/science.1247005
6. H. Koike-Yusa, Y. Li, E. P. Tan, M. C. Velasco-Herrera, K. Yusa, Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2014). Medline doi:10.1038/nbt.2800
7. J. E. Carette, C. P. Guimaraes, M. Varadarajan, A. S. Park, I. Wuethrich, A. Godarova, M. Kotecki, B. H. Cochran, E. Spooner, H. L. Ploegh, T. R. Brummelkamp, Haploid genetic screens in human cells identify host factors used by pathogens. *Science* **326**, 1231–1235 (2009). Medline doi:10.1126/science.1178955
8. J. E. Carette, M. Raaben, A. C. Wong, A. S. Herbert, G. Obernosterer, N. Mulherkar, A. I. Kuehne, P. J. Kranzusch, A. M. Griffin, G. Ruthel, P. Dal Cin, J. M. Dye, S. P. Whelan, K. Chandran, T. R. Brummelkamp, Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* **477**, 340–343 (2011). Medline doi:10.1038/nature10348
9. I. A. Tchasovnikarova, R. T. Timms, N. J. Matheson, K. Wals, R. Antrobus, B. Göttgens, G. Dougan, M. A. Dawson, P. J. Lehner, Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. *Science* **348**, 1481–1485 (2015). Medline doi:10.1126/science.aaa7227
10. J. P. Kastenmayer, L. Ni, A. Chu, L. E. Kitchen, W. C. Au, H. Yang, C. D. Carter, D. Wheeler, R. W. Davis, J. D. Boeke, M. A. Snyder, M. A. Basrai, Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res.* **16**, 365–373 (2006). Medline doi:10.1101/gr.4355406
11. G. S. Cowley *et al*., Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data* 10.1038/sdata.2014.35 (2014).
12. A. C. Wilson, S. S. Carlson, T. J. White, Biochemical evolution. *Annu. Rev. Biochem.* **46**, 573–639 (1977). Medline doi:10.1146/annurev.bi.46.070177.003041
13. Y. Ishihama, T. Schmidt, J. Rappsilber, M. Mann, F. U. Hartl, M. J. Kerner, D. Frishman, Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* **9**, 102 (2008). Medline doi:10.1186/1471-2164-9-102
14. H. Jeong, S. P. Mason, A. L. Barabási, Z. N. Oltvai, Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001). Medline doi:10.1038/35075138
15. T. Hart, K. R. Brown, F. Sircoulomb, R. Rottapel, J. Moffat, Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Sys. Biol.* 10.15252/msb.20145216 (2014).
16. Z. Gu, L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, W. H. Li, Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66 (2003). Medline doi:10.1038/nature01198
17. H. Liang, W.-H. Li, Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* **23**, 375–378 (2007). Medline doi:10.1016/j.tig.2007.04.005
18. B.-Y. Liao, J. Zhang, Mouse duplicate genes are as essential as singletons. *Trends Genet.* **23**, 378–381 (2007). Medline doi:10.1016/j.tig.2007.05.006
19. T. Makino, K. Hokamp, A. McLysaght, The complex relationship of gene duplication and essentiality. *Trends Genet.* **25**, 152–155 (2009). Medline doi:10.1016/j.tig.2009.03.001
20. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005). Medline doi:10.1073/pnas.0506580102
21. Y. Ahmad, F.-M. Boisvert, E. Lundberg, M. Uhlen, A. I. Lamond, Systematic analysis of protein pools, isoforms, and modifications affecting turnover and subcellular localization. *Mol. Cell. Proteomics* **11**, 013680 (2012). 10.1074/mcp.M111.013680 Medline doi:10.1074/mcp.M111.013680
22. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi Jr., M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, L. A. Garraway, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012). Medline doi:10.1038/nature11003
23. A. G. Baltz, M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, M. Landthaler, The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012). Medline doi:10.1016/j.molcel.2012.05.021
24. T. Sekiguchi, H. Iida, J. Fukumura, T. Nishimoto, Human DDX3Y, the Y-encoded isoform of RNA helicase DDX3, rescues a hamster temperature-sensitive ET24 mutant cell line with a DDX3X mutation. *Exp. Cell Res.* **300**, 213–222 (2004). Medline doi:10.1016/j.yexcr.2004.07.005
25. F. L. Muller, S. Colla, E. Aquilanti, V. E. Manzo, G. Genovese, J. Lee, D. Eisenson, R. Narurkar, P. Deng, L. Nezi, M. A. Lee, B. Hu, J. Hu, E. Sahin, D. Ong, E. Fletcher-Sananikone, D. Ho, L. Kwong, C. Brennan, Y. A. Wang, L. Chin, R. A. DePinho, Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature* **488**, 337–342 (2012). Medline doi:10.1038/nature11331
26. K. Ohneda, M. Yamamoto, Roles of hematopoietic transcription factors GATA-1 and GATA-2 in the development of red blood cell lineage. *Acta Haematol.* **108**, 237–245 (2002). Medline doi:10.1159/000065660
27. L. C. Andersson, K. Nilsson, C. G. Gahmberg, K562—a human erythroleukemic

cell line. *Int. J. Cancer* **23**, 143–147 (1979). Medline doi:10.1002/ijc.2910230202

28. B. S. Andersson *et al.*, KBM-7, a human myeloid leukemia cell line with double Philadelphia chromosomes lacking normal c-ABL and BCR transcripts. *Leukemia* **9**, 2100–2108 (1995). Medline

29. S. Q. Wu, K. V. Voelkerding, L. Sabatini, X. R. Chen, J. Huang, L. F. Meisner, Extensive amplification of bcr/abl fusion genes clustered on three marker chromosomes in human leukemic cell line K-562. *Leukemia* **9**, 858–862 (1995). Medline

30. A. Constantinou, K. Kiguchi, E. Huberman, Induction of differentiation and DNA strand breakage in human HL-60 and K-562 leukemia cells by genistein. *Cancer Res.* **50**, 2618–2624 (1990). Medline

31. B. Scappini, S. Gatto, F. Onida, C. Ricci, V. Divoky, W. G. Wierda, M. Andreeff, L. Dong, K. Hayes, H. Verstovsek, H. M. Kantarjian, M. Beran, Changes associated with the development of resistance to imatinib (STI571) in two leukemia cell lines expressing p210 Bcr/Abl protein. *Cancer* **100**, 1459–1471 (2004). Medline doi:10.1002/cncr.20131

32. S. Iida, P. H. Rao, P. Nallasivam, H. Hibshoosh, M. Butler, D. C. Louie, V. Dyomin, H. Ohno, R. S. Chaganti, R. Dalla-Favera, The t(9;14)(p13;q32) chromosomal translocation associated with lymphoplasmacytoid lymphoma involves the PAX-5 gene. *Blood* **88**, 4110–4117 (1996). Medline

33. H. Bouamar, S. Abbas, A. P. Lin, L. Wang, D. Jiang, K. N. Holder, M. C. Kinney, S. Hunicke-Smith, R. C. Aguiar, A capture-sequencing strategy identifies IRF8, EBF1, and APRIL as novel IGH fusion partners in B-cell lymphoma. *Blood* **122**, 726–733 (2013). Medline doi:10.1182/blood-2013-04-495804

34. S. Galiègue-Zouitina, S. Quief, M. P. Hildebrand, C. Denis, G. Lecocq, M. Collyn-d'Hooghe, C. Bastard, M. Yuille, M. J. Dyer, J. P. Kerckaert, Fusion of the LAZ3/BCL6 and BOB1/OBF1 genes by t(3; 11) (q27; q23) chromosomal translocation. *C. R. Acad. Sci. III* **318**, 1125–1131 (1995). Medline

35. B. Chapuy, M. R. McKeown, C. Y. Lin, S. Monti, M. G. Roemer, J. Qi, P. B. Rahl, H. H. Sun, K. T. Yeda, J. G. Doench, E. Reichert, A. L. Kung, S. J. Rodig, R. A. Young, M. A. Shipp, J. E. Bradner, Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer Cell* **24**, 777–790 (2013). Medline doi:10.1016/j.ccr.2013.11.003

36. R. D. Morin, M. Mendez-Lago, A. J. Mungall, R. Goya, K. L. Mungall, R. D. Corbett, N. A. Johnson, T. M. Severson, R. Chiu, M. Field, S. Jackman, M. Krzywinski, D. W. Scott, D. L. Trinh, J. Tamura-Wells, S. Li, M. R. Firme, S. Rogic, M. Griffith, S. Chan, O. Yakovenko, I. M. Meyer, E. Y. Zhao, D. Smailus, M. Moksa, S. Chittaranjan, L. Rimsza, A. Brooks-Wilson, J. J. Spinelli, S. Ben-Neriah, B. Meissner, B. Woolcock, M. Boyle, H. McDonald, A. Tam, Y. Zhao, A. Delaney, T. Zeng, K. Tse, Y. Butterfield, I. Birol, R. Holt, J. Schein, D. E. Horsman, R. Moore, S. J. Jones, J. M. Connors, M. Hirst, R. D. Gascoyne, M. A. Marra, Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298–303 (2011). Medline doi:10.1038/nature10351

37. J. M. Engreitz, A. Pandya-Jones, P. McDonel, A. Shishkin, K. Sirokman, C. Surka, S. Kadri, J. Xing, A. Goren, E. S. Lander, K. Plath, M. Guttman, The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973 (2013). Medline doi:10.1126/science.1237973

38. D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013). Medline doi:10.1186/gb-2013-14-4-r36

39. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010). Medline doi:10.1038/nbt.1621

40. F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, P. Flicek, Ensembl 2015. *Nucleic Acids Res.* **43** (D1), D662–D669 (2015). Medline doi:10.1093/nar/gku1010

41. L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, S. H. Bryant, The NCBI BioSystems database. *Nucleic Acids Res.* **38** (Database),

D492–D496 (2010). Medline doi:10.1093/nar/gkp858

42. P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, S. M. Searle, Ensembl 2014. *Nucleic Acids Res.* **42** (D1), D749–D755 (2014). Medline doi:10.1093/nar/gkt1196

43. J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey Broad GO Seattle GO NHLBI Exome Sequencing Project, Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012). Medline doi:10.1126/science.1219240

44. C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006). Medline doi:10.1093/nar/gkj109

45. H. Li, A. Coghlan, J. Ruan, L. J. Coin, J. K. Hériché, L. Osmotherly, R. Li, T. Liu, Z. Zhang, L. Bolund, G. K. Wong, W. Zheng, P. Dehal, J. Wang, R. Durbin, TreeFam: A curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580 (2006). Medline doi:10.1093/nar/gkj118

46. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi Jr., M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, L. A. Garraway, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012). Medline doi:10.1038/nature11003

47. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009). Medline doi:10.1038/nprot.2008.211

48. P. Mertins, N. D. Udeshi, K. R. Clauser, D. R. Mani, J. Patel, S. E. Ong, J. D. Jaffe, S. A. Carr, iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. *Mol. Cell. Proteomics* **11**, 014423 (2012). 10.1074/mcp.M111.014423 Medline doi:10.1074/mcp.M111.014423

49. S. Schwartz, M. R. Mumbach, M. Jovanovic, T. Wang, K. Maciag, G. G. Bushkin, P. Mertins, D. Ter-Ovanesyan, N. Habib, D. Cacchiarelli, N. E. Sanjana, E. Freinkman, M. E. Pacold, R. Satija, T. S. Mikkelsen, N. Hacohen, F. Zhang, S. A. Carr, E. S. Lander, A. Regev, Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. *Cell Reports* **8**, 284–296 (2014). Medline doi:10.1016/j.celrep.2014.05.048

50. A. G. Uren, H. Mikkers, J. Kool, L. van der Weyden, A. H. Lund, C. H. Wilson, R. Rance, J. Jonkers, M. van Lohuizen, A. Berns, D. J. Adams, A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat. Protoc.* **4**, 789–798 (2009). Medline doi:10.1038/nprot.2009.64

51. T. Brady, Y. N. Lee, K. Ronen, N. Malani, C. C. Berry, P. D. Bieniasz, F. D. Bushman, Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* **23**, 633–642 (2009). Medline doi:10.1101/gad.1762309

## ACKNOWLEDGMENTS

**SUPPLEMENTARY MATERIALS**

www.sciencemag.org/cgi/content/full/science.aac7041/DC1
Materials and Methods
Supplementary Text S1 to S5
Figs. S1 to S6
Tables S1 to S6
References (*37–51*)
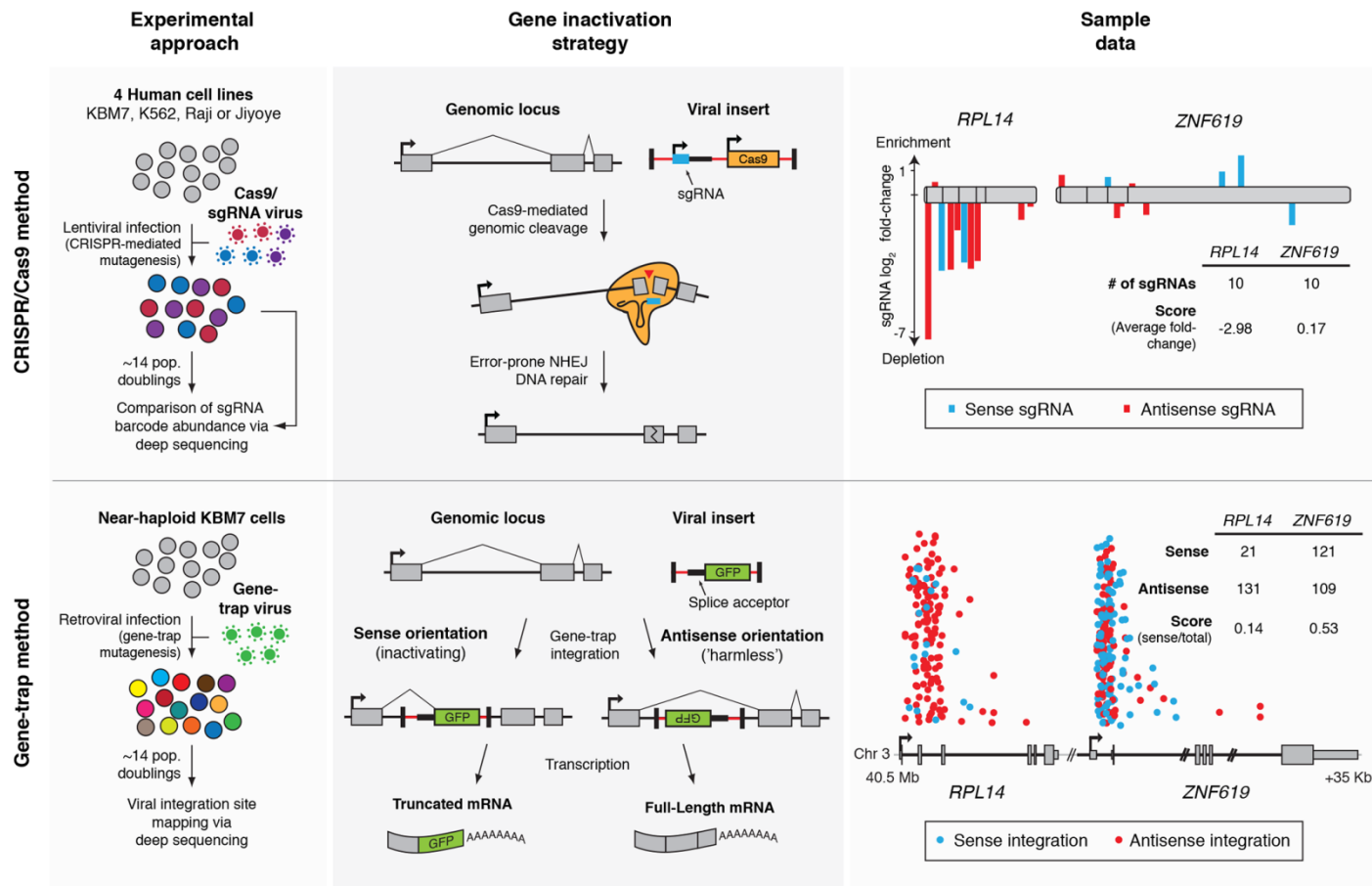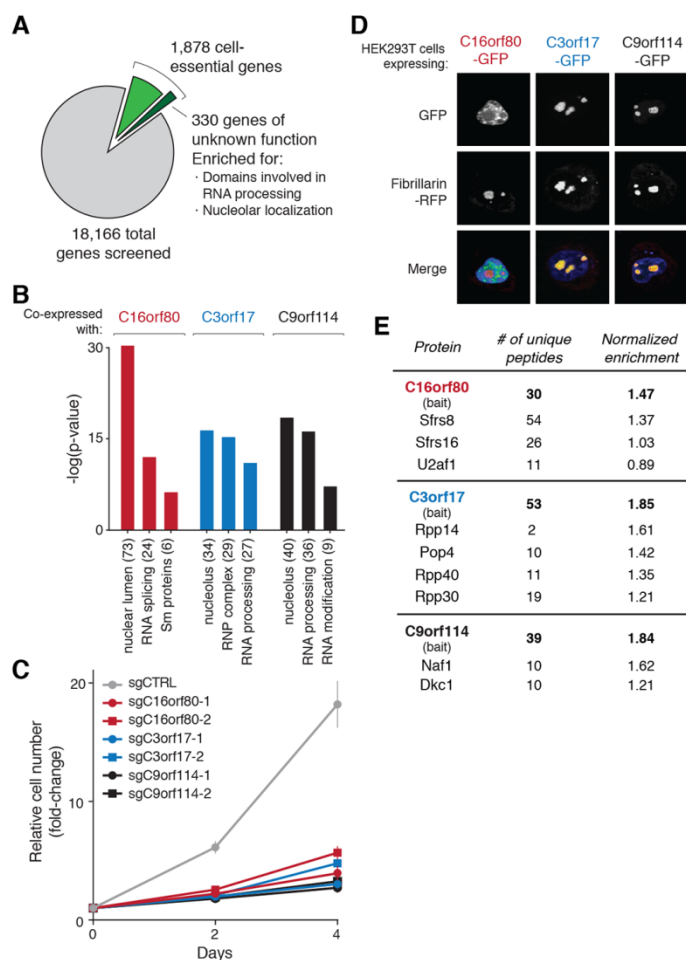
**Fig. 1. Two approaches for genetic screening in human cells.** (**Top**) CRISPR/Cas9. Cells are transduced with a genome-wide sgRNA lentiviral library. Gene inactivation via Cas9-mediated genomic cleavage is directed by the 20-bp sequence at the 5′ end of the sgRNA. Cells bearing sgRNAs targeting essential genes are depleted in the final population. (**Bottom**) Gene-trap. KBM7 cells are transduced with a gene-trap retrovirus which integrates in an inactivating or "harmless" orientation at random genomic loci. Essential genes contain fewer insertions in the inactivating orientation. Sample data for two neighboring genes *RPL14*, encoding an essential ribosomal protein, and *ZNF619*, encoding a dispensable zinc finger protein, are displayed. For CRISPR/Cas9, sgRNAs are plotted according to their target position along each gene with the height of each bar indicating the level of depletion. Boxes indicate individual exons. For gene-trap, the intronic insertion sites in each gene are plotted according to their orientation and genomic position. The height of each point is randomized.
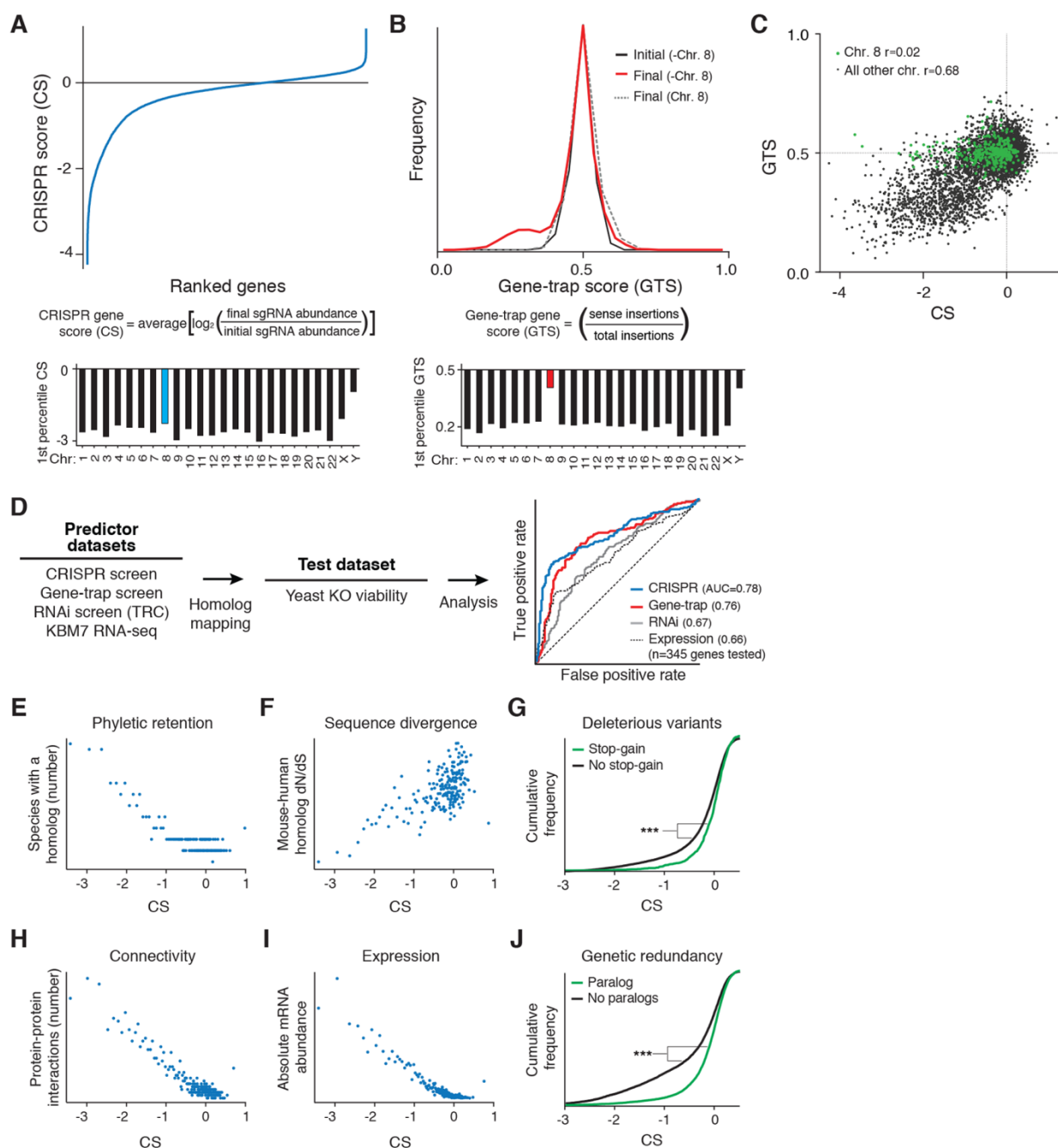
**Fig. 2. Identification and characterization of human cell-essential genes.** (**A**) CRISPR scores (CS) of all genes in the KBM7 cells. Similar proportions of cell-essential genes were identified on all autosomes. (**B**) KBM7 gene-trap score (GTS) distributions. No low GTS genes were detected on the diploid chromosome 8. (**C**) CS and GTS of overlapping genes. (**D**) Yeast homolog essentiality prediction analysis. (**E**) Broader retention of essential genes across species. (**F**) Higher sequence conservation of essential genes. (**G**) Genes with deleterious stop-gain variants are less likely to be essential. (**H**) Greater connectivity of proteins encoded by essential genes. (**I**) Higher mRNA transcript levels of essential genes. (**J**) Genes with paralogs are less likely to be essential. ***$P < 0.001$ from Kolmogorov-Smirnov test.

**A**

1,878 cell-essential genes

330 genes of unknown function
Enriched for:
· Domains involved in RNA processing
· Nucleolar localization

18,166 total genes screened

**B**

Co-expressed with:

C16orf80    C3orf17    C9orf114

-log(p-value)

nuclear lumen (73)
RNA splicing (24)
Sm proteins (6)
nucleolus (34)
RNP complex (29)
RNA processing (27)
nucleolus (40)
RNA processing (36)
RNA modification (9)

**C**

Relative cell number (fold-change)

sgCTRL
sgC16orf80-1
sgC16orf80-2
sgC3orf17-1
sgC3orf17-2
sgC9orf114-1
sgC9orf114-2

Days

**D**

HEK293T cells expressing:

C16orf80-GFP    C3orf17-GFP    C9orf114-GFP

GFP

Fibrillarin-RFP

Merge

**E**

| Protein | # of unique peptides | Normalized enrichment |
|---|---|---|
| **C16orf80** (bait) | **30** | **1.47** |
| Sfrs8 | 54 | 1.37 |
| Sfrs16 | 26 | 1.03 |
| U2af1 | 11 | 0.89 |
| **C3orf17** (bait) | **53** | **1.85** |
| Rpp14 | 2 | 1.61 |
| Pop4 | 10 | 1.42 |
| Rpp40 | 11 | 1.35 |
| Rpp30 | 19 | 1.21 |
| **C9orf114** (bait) | **39** | **1.84** |
| Naf1 | 10 | 1.62 |
| Dkc1 | 10 | 1.21 |

**Fig. 3. Functional characterization of novel cell-essential genes.** (**A**) Of the 1,878 cell-essential genes identified in the KBM7 cell line, 330 genes had no known function. (**B**) Genes co-expressed with *C16orf80*, *C3orf17*, and *C9orf114* across CCLE cell lines were associated with RNA processing. Parentheses denote the number of genes in the set. (**C**) Proliferation of KBM7 cells transduced with sgRNAs targeting *C16orf80*, *C3orf17*, and *C9orf114*, or a non-targeting control. Error bars denote SD ($n = 4$). (**D**) C16orf80 localized to the nucleus and C3orf17 and C9orf114 to the nucleolus. Fibrillarin-RFP was used as a nucleolar marker. (**E**) Multiple subunits of the spliceosome, RNAse P/MRP and H/ACA ribonucleoprotein complexes interact with C16orf80, C3orf17, and C9orf114, respectively.
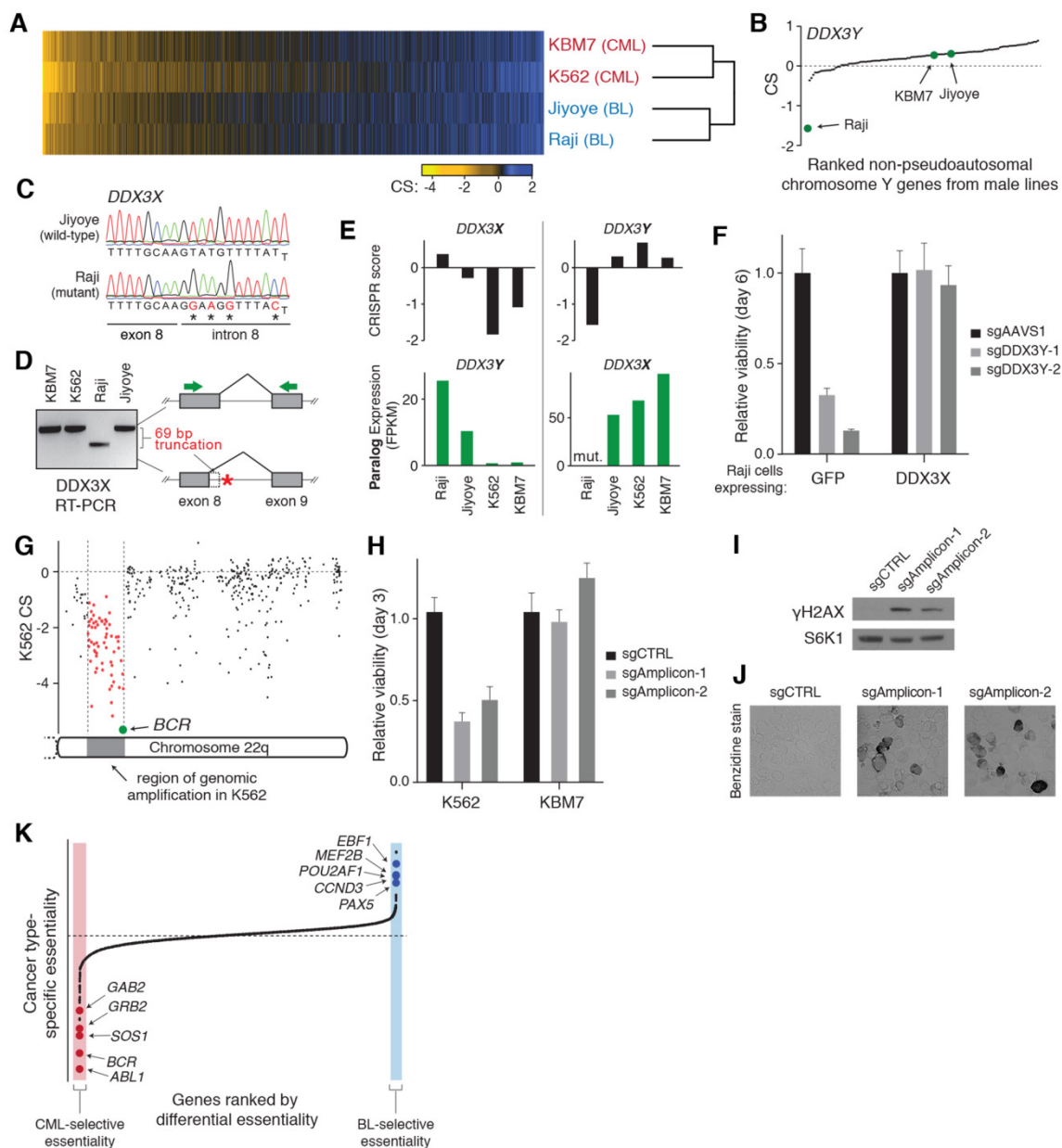
**Fig. 4. Comparisons of gene essentiality across four cell lines.** (**A**) Heatmap of CS across four cell lines sorted by average CS. (**B**) CS of genes residing in the non-pseudoautosomal region of chromosome Y for male cell lines (Raji, Jiyoye and KBM7). (**C**) Sanger sequencing of *DDX3X* in Raji cells reveals mutations in the 5′ splice site of intron 8. (**D**) Splice-site mutations in *DDX3X* result in a 69-bp truncation of the mRNA. PCR primers spanning exons 8 and 9 of *DDX3X* (denoted by green arrows) were used to amplify cDNA from each line. (**E**) Essentiality of *DDX3X* and *DDX3Y* is determined by the expression and mutation status of their paralogs. (**F**) Proliferation of GFP- and DDX3X-expressing Raji cells transduced with sgRNAs targeting *DDX3Y* or an AAVS1-targeting control. Error bars denote SD ($n = 4$). (**G**) Analysis of genes on chromosome 22q reveals apparent 'essentiality' of 61 contiguous genes in K562 residing in a region of high-copy tandem amplification. (**H**) Proliferation of K562 and KBM7 cells transduced with sgRNAs targeting a non-genic region within the BCR-ABL amplicon in K562 cells or a non-targeting control. Error bars denote SD ($n = 4$). (**I**) γH2AX (phospho-S139 H2AX) immunoblot analysis of K562 transduced with sgRNAs as in (H). S6K1 was used as a loading control. (**J**) Cleavage within amplified region induced erythroid differentiation in K562 cells as assessed by 3,3′-dimethoxybenzidine hemoglobin staining. (**K**) Comparison of gene essentiality between the two cancer types reveals oncogenic drivers and lineage specifiers. Genes were ranked by the difference between the average CS of each cancer type.

# Science

## Identification and characterization of essential genes in the human genome

Tim Wang, Kivanç Birsoy, Nicholas W. Hughes, Kevin M. Krupczak, Yorick Post, Jenny J. Wei, Eric S. Lander and David M. Sabatini

| | |
|---|---|
| **ARTICLE TOOLS** | http://science.sciencemag.org/content/early/2015/10/14/science.aac7041 |
| **SUPPLEMENTARY MATERIALS** | http://science.sciencemag.org/content/suppl/2015/10/14/science.aac7041.DC1 |
| **RELATED CONTENT** | http://science.sciencemag.org/content/sci/350/6264/1028.full<br>http://science.sciencemag.org/content/sci/350/6264/1092.full |
| **REFERENCES** | This article cites 51 articles, 19 of which you can access for free<br>http://science.sciencemag.org/content/early/2015/10/14/science.aac7041#BIBL |
| **PERMISSIONS** | http://www.sciencemag.org/help/reprints-and-permissions |

Use of this article is subject to the Terms of Service